Supplementary Material – *Intervening in Black Box: Concept Bottleneck Model for Enhancing Human Neural Network Mutual Understanding*

Nuoye Xiong[1], Anqi Dong[2*], Ning Wang[1], Cong Hua[1], Guangming Zhu[1], Lin Mei[3], Peiyi Shen[1], Liang Zhang[1†]
[1]Xidian University, China
[2]KTH Royal Institute of Technology, Sweden
[3]Donghai Laboratory, China
{nyx, ningwang, chua}@stu.xidian.edu.cn
{gmzhu, pyshen, liangzhang}@xidian.edu.cn
{meilin}@donghailab.com
{anqid}@kth.se

Outline

The organization of supplementary material is as follows: Section A introduces the concept replacement method based on additional concept search sets. Section B illustrates the optimization of predictions before and after intervention within an interpretable CBM, along with an analysis of its approximation to a black-box model for generalization. Also, the change in black-box classification accuracy for non-intervened classes before and after the intervention is reported. Section C details an intervention-based concept masking experiment for vision-related tasks. Sections D through G provide further experimental details and discussions. Finally, Section H presents additional visualizations of intervention explanations. To address the concept bottleneck limitations.

A. Concept Replacement

It is also noteworthy that simply deleting concepts may not resolve classification errors caused by conceptual bottleneck limitations. To address this, we propose a concept replacement method that leverages a search set of additional concepts.

The framework of the replacement method is illustrated in Figure I. It begins by identifying concepts that require intervention within each confusion class through concept intervention. These concepts are then ranked based on their frequency of occurrence, and the top $\bar{q}$ most frequently occurring concepts are selected for replacement. For each identified concept, positive concepts that do not require modification are also determined. Simultaneously, a replacement concept is selected from an external search set, with the objective of positioning it as far as possible from the embedding of negative concepts while bringing it closer to positive concepts. Cosine similarity is utilized to measure expression similarity, allowing us to score potential replacement concepts effectively.

Figure II presents the accuracy comparison of NFResNet50 on Flower-102 before and after concept replacement. The results demonstrate that replacing different numbers of concepts leads to improved test accuracy for both the black-box baseline and the corresponding CBM.

B. More Evaluation on Approximation and Intervention

We also conduct a class-based accuracy analysis on the approximate black-box CBM inference structure and compare it with the original black-box model. Figure III presents a subset of accuracy comparison curves for confusing classes between the CBM-based approximate reasoning structure and the original black box. Additionally, Table I reports improvements achieved by CBM inference after applying a

*Co-first contributing author.
†Corresponding author.

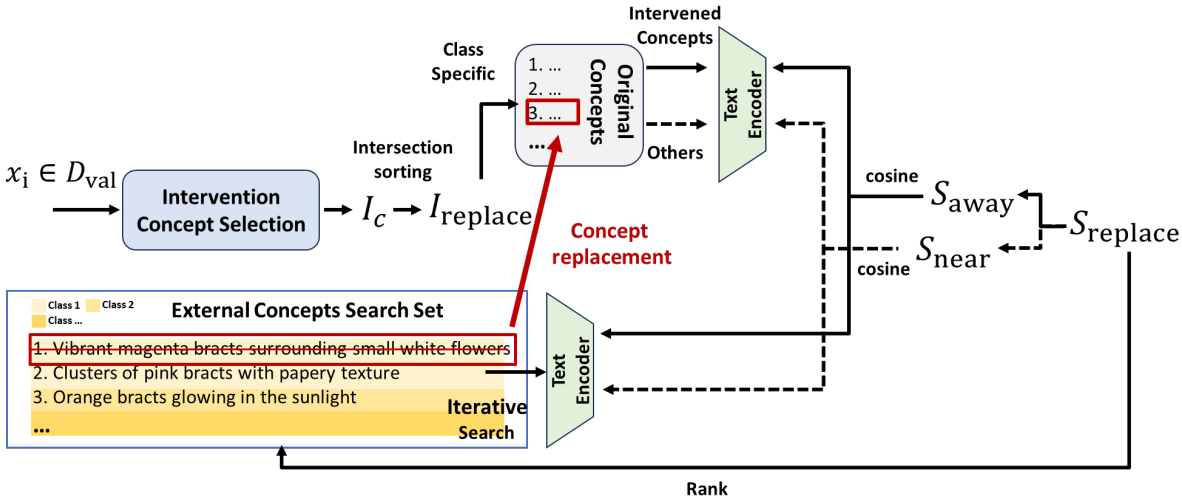

Figure I. Concepts replacement framework. **Intervention Concept Selection:** Obtain the concepts that require intervention for each confused class. **Intersection Sorting:** The intervening concepts for all confused classes are intersected, and the concepts that need to be replaced are obtained by sorting the number of occurrences. **Concept Replacement:** Iteratively search for candidates in the External Concepts Search Set that are far away from the replacement concept but similar to other positive concepts of the same class, sort according to the $S_{\mathbf{replace}} = S_{\mathbf{near}} - S_{\mathbf{away}}$, and select the optimal concept replacement.

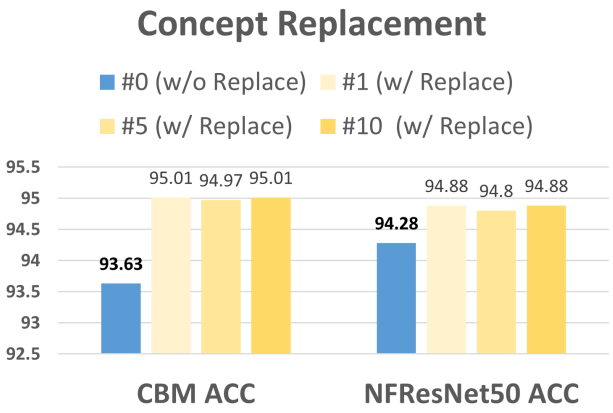

Figure II. Comparison of accuracy before and after the concept replacement of NFResNet50 and the corresponding CBM on Flower-102.

concept weight matrix in CBM-HNMU. The results indicate that the classification accuracy curve of CBM on intervention classes generally aligns well with that of the original black box. This suggests that CBM can effectively approximate the reasoning logic of the black box within locally confused classes. Furthermore, we evaluate the accuracy of the concept intervention algorithm based on CBM. The findings reveal that nearly all interventions help correct concept-related reasoning errors, leading to improvements in both class-level and global accuracy.

To preserve the predicted distribution over non-target classes when transferring intervention knowledge, we apply the operator – $\mathbf{pr}$, to reassign only the residual probability mass from non-target classes to the intervention class, as defined in Eq. (9). Figure IV reports the accuracy shift for non-target classes across both models and all three datasets. The results show that performance is stable and closely aligned with pre-intervention accuracy (scatter plot), with a slight overall improvement (box plot).

C. Intervention Concept Visual Masking

We use OpenAI-CLIP as the concept communication module, enabling CBM-HNMU to align natural language intervention semantics with corresponding visual concept changes before and after intervention. To validate this alignment, we conduct visual masking experiments on intervention-related concepts across different models and datasets.

First, we identify a specific class of intervention concepts in the black-box model and extract the top-ranked natural language concepts with the highest intervention scores ($S_{\mathrm{nT}}/S_{\mathrm{pF}}$) that are also visually relevant. Next, we collect all samples from that class that the black-box model misclassifies. Finally, we apply pixel-based masking to mask the corresponding semantic information associated with the intervening natural language concept, generating new sample inputs.

These modified samples are then fed into the original black-box model to obtain new predictions and confidence scores, and compare with the original results. As shown in Figure V, in the first example, we visually masked 7 misclassifies samples of *sword lily* on Flower-102 using NFResNet-50. Based on the vision-related concepts identified in the intervention's natural language description, we masked the central pixels of the flowers in these samples.

Models	Flower-102		CUB-200		FGVC-Aircraft	
	$w/o\ INT$	$w/\ INT$	$w/o\ INT$	$w/\ INT$	$w/o\ INT$	$w/\ INT$
NFResNet50	93.63	93.75 (↑ 0.12)	61.90	62.70 (↑ 0.80)	65.80	66.67 (↑ 0.87)
Vit_Small	80.10	80.28 (↑ 0.18)	54.10	54.15 (↑ 0.05)	62.95	63.55 (↑ 0.60)
ResNeXt26	89.16	90.69 (↑ 1.53)	60.15	60.50 (↑ 0.35)	63.79	64.69 (↑ 0.90)
GCVit_Base	92.84	93.26 (↑ 0.42)	77.10	77.15 (↑ 0.05)	68.98	69.91 (↑ 0.93)

Table I. Comparison of classification accuracy with P_{CBM} $(w/o\ INT)$ and $\tilde{P}_{CBM}$ (w/INT).

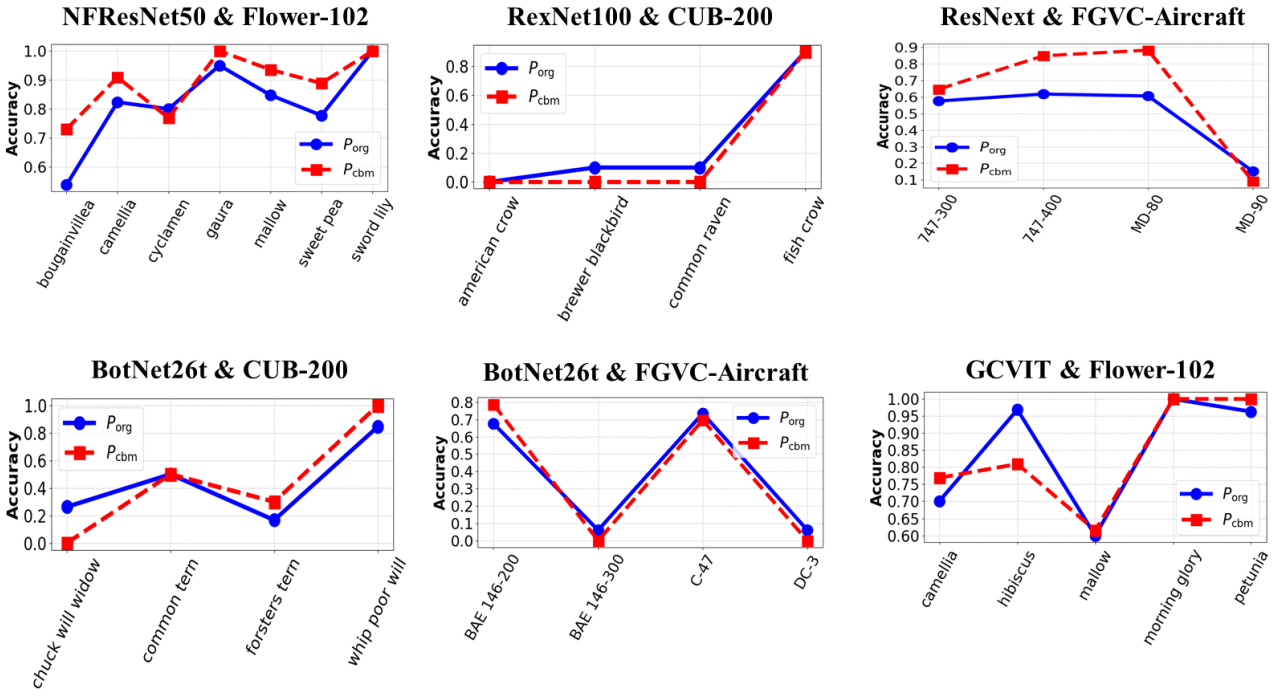

Figure III. Confusion class accuracy comparison curve between part of the CBM approximate reasoning structure and the original black box. The y-axis is the prediction accuracy of the class, and the x-axis is the corresponding intervention class (Γ).

In the second example, we performed a similar procedure on four misclassified samples spanning three aircraft classes (*747-300, 747-400 and DC-3*). Notably, when conducting visual masking experiments on CUB-200 using GCVit, the intervention concept exhibited the semantic feature of gray. To account for this, we applied a pure white mask to cover the pixels corresponding to the intervention concept.

D. Hardware and Software Settings

All experiments in this work are conducted on Ubuntu 20.04 within an Anaconda3 virtual environment, using NVIDIA 3090 (24GB) GPU. This setup allows us to provide integrated environment resources, including code and datasets, in a public remote hub in the future.

The pre-trained network weights and datasets referenced in the paper are publicly available resources. We will include download links and deployment instructions in subsequent packaged code. Additionally, for the various methods cited in the paper, we will provide links along with detailed deployment guidelines.

E. Parameter Settings

All baseline models ($P_{org}\ w/o\ INT$) use weights pre-trained on ImageNet-1K, and are tuned on the corresponding experimental datasets with 50 epoch. During the fine-tuning process, the learning rate is set to $1e^{-}4$. Both confusing classes selection, local approximation are performed on the corresponding D_{val}. During local approximation, learning rate is set to $1e^{-}4$ and epoch is set to 200. Concepts intervention is still performed on the D_{val} and only execute the Algorithm 1 in the paper once to modify the concept weight matrix (W) of the corresponding P_{CBM}.

Knowledge transfer requires setting different distillation

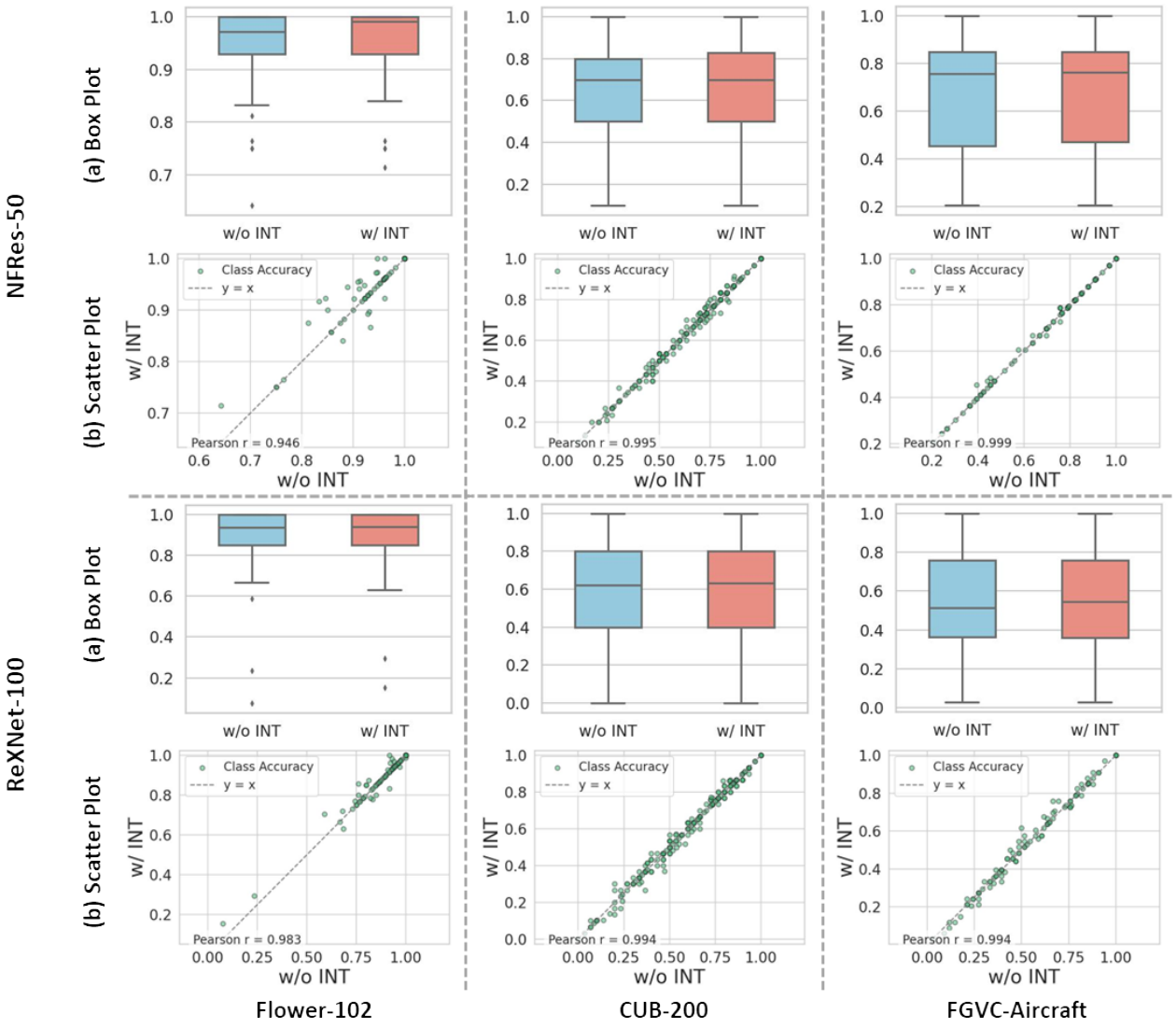

Figure IV. Before and after intervention in non-intervention classes.

temperatures for teachers and student, where the teacher model P_t is spliced by the frozen original black box (P_t^1) and locally approximated CBM (P_t^2). The student model P_s is the original black box. The distillation temperature T_1 of P_t^2 is 2.0, and the distillation temperature T_2 of P_s is 1.5. Knowledge transfer is performed on the D_{val} for 10 epoch with learning rate $3e^{-}7$. The maximum number of intervention concepts varies depending on the datasets. It is recommended to set it between 10 and 100. The number can be adjusted according to the intervention effect (we take the optimal value in multiple groups of experiments).

F. Discussion

In this paper, we demonstrate the effectiveness of the CBM-HNMU approach combined with gradient-based intervention. In fact, according to the human-understandable intervention concepts provided by CBM-HNMU, we can even manually select the visual part corresponding to the concept to quickly intervene and determine the harmful concept dependence of the model on the data domain. Secondly, CBM-HNMU bridges the black box and interpretable structures, integrates visual and language modalities, and provides intuitive model explanations for easy understanding. The method's explanation-based intervention effectively identifies the recognition patterns and biases inherent in black-box models, laying the foundation for building interpretable classification networks in the future.

G. Limitations

CBM-HNMU relies on both visual concepts and natural language concepts extracted from the black box. Although many unsupervised methods can be used to efficiently ex-

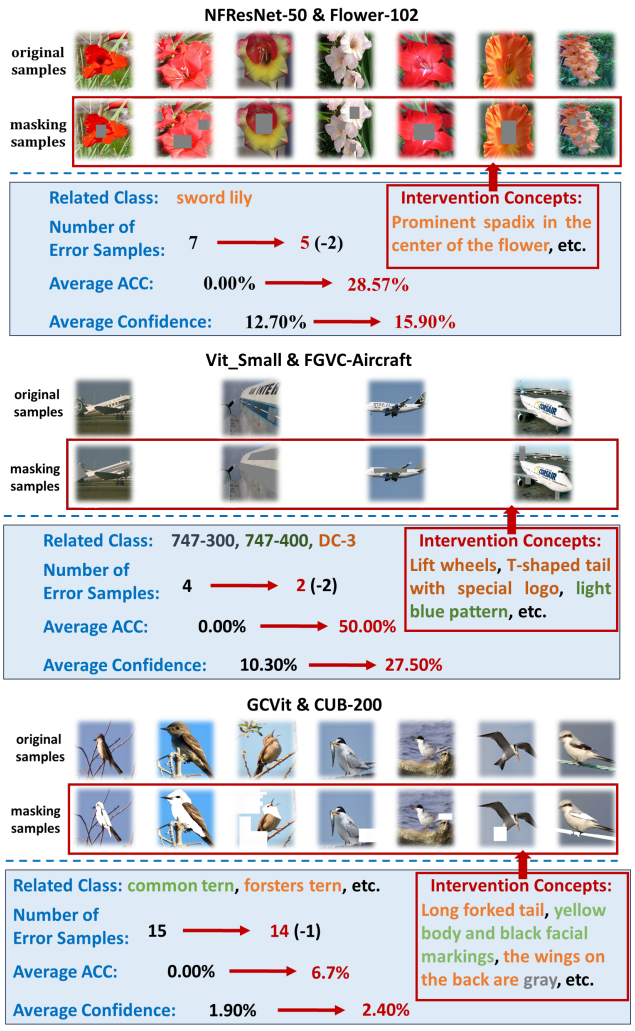

Figure V. Visual related concepts masking. Intervention concepts and class name having the same color represent that the intervention concept belongs to the corresponding class, and the visual mask of samples belonging to this class is labeled according to these concepts.

tract the concepts of the corresponding model and even give attribution explanations, and using LLMs can quickly obtain the concept bottleneck of the corresponding datasets, inevitably due to 1) limitations of concept extraction methods, such as concept extraction and explanation methods that are not model-oriented, it will lead to situations where concepts cannot be well connected to samples. 2) The hallucination phenomenon of LLMs may produce a large number of abstract concepts. Abstract concepts and concrete concepts have little impact on the expression of concept bottleneck and OpenAI-CLIP can be also used to connect abstract concepts with visual feature. However, it is disadvantageous for human to understand the model intervention process and error correction explanation based on the relationship between abstract concepts and visual concepts. Problem 1) can be solved by applying model-oriented concept extraction methods, such as the *Model-Oriented Concept Extraction* (MOCE). For question 2), manual verification is a more compromised method, which can save most of the time while ensuring the quality of the concept.

H. Visualization and Explanation

We begin with additional visualizations of intervention-based explanations, as shown in Figures VI – XIV. These illustrations highlight the relationship between natural language intervention concepts and changes in black-box visual representations before and after intervention. We present results using CBM-NHMU on NFResNet50, BotNet26, and RexNet100, with interventions applied to Flower-102, CUB-200, and FGVC-Aircraft. Notably, the samples are randomly selected from a subset where the black-box model's original classification errors on the test set are corrected after intervention. Each corrected sample includes at least one pre- or post-correction class associated with the confused classes, ensuring a clear visual link to the intervention concept (see "Coverage" in the manuscript).

Before detailing each intervention explanation example, we first clarify the key components in each visualization. In each example, the upper-left image represents the CRAFT concept attribution of the input image to the post-intervention black-box model. Recall P_{org} and $P_{w/INT}$ denote the classification predictions of the black-box model before and after the intervention, respectively.

Incorrect class predictions (i.e., those made by the original black-box model) are marked **green** if they belong to the confused classes (Γ) and **black** otherwise. The correct class prediction (i.e., the black-box model's output after intervention) is marked **purple** if it falls within the confusion category; otherwise, it is also marked by **black**. In each visualization below, real example images corresponding to incorrectly predicted classes are displayed. These examples help illustrate why the black-box model misclassified the input and how the intervention concept corrects the error. By examining images from confused categories alongside intervention concepts, users can better understand the reasoning behind the intervention. Intervention concepts and their semantically related visual counterparts in the confused classes are highlighted with **purple outlines**, whereas relevant concepts from the incorrectly predicted class are enclosed in **green borders**. We are now in the position to detail intervention explanation examples.

In Figure VI, the baseline of NFResNet50 misclassifies *sword lily* as *bougainvillea*. The visually related intervention concept: *"Prominent spadix in the center of the flower."* prompts us to delete the related concept about the prominence in the center of flower to correct this error. We can find that the third most important visual concept extracted by the original black box before intervention corresponds

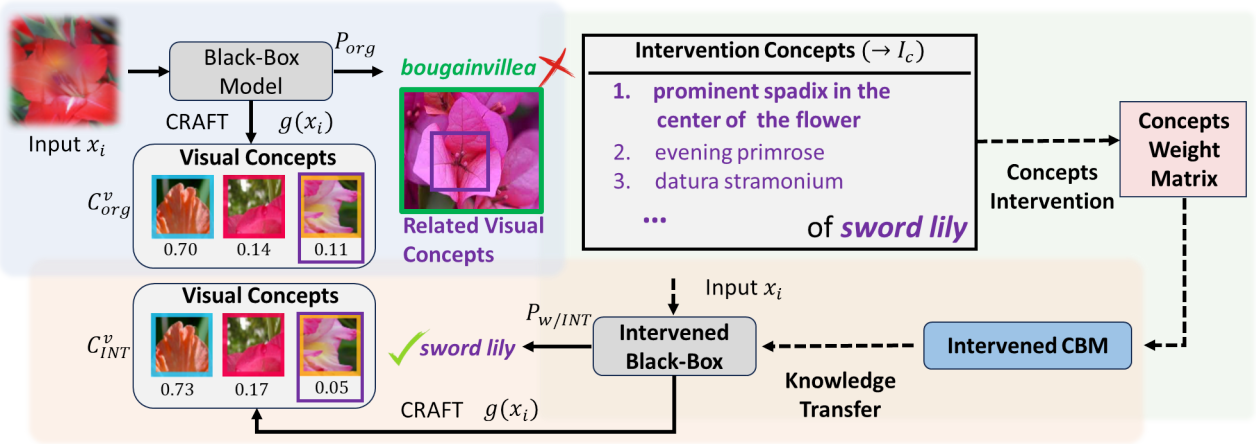

Figure VI. More visualization of NFResNet50 on Flower-102.

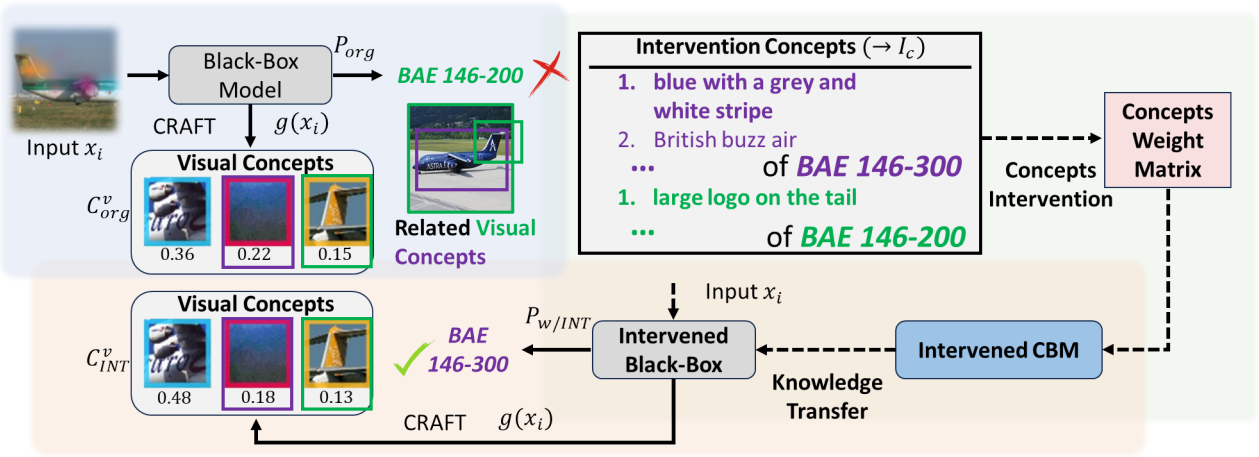

Figure VII. More visualization of NFResNet50 on FGVC-Aircraft.

exactly to this description, and the importance score of this concept is 0.11. After the intervention, the black box gives a similar conceptual explanation, but the difference is that the importance score of the intervention concept dropped

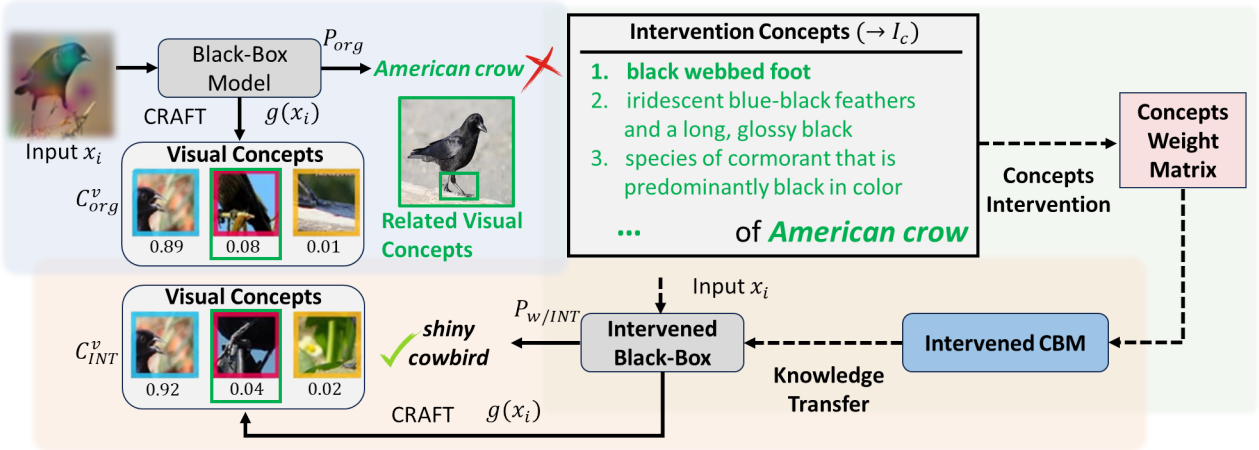

Figure VIII. More visualization of NFResNet50 on CUB-200.

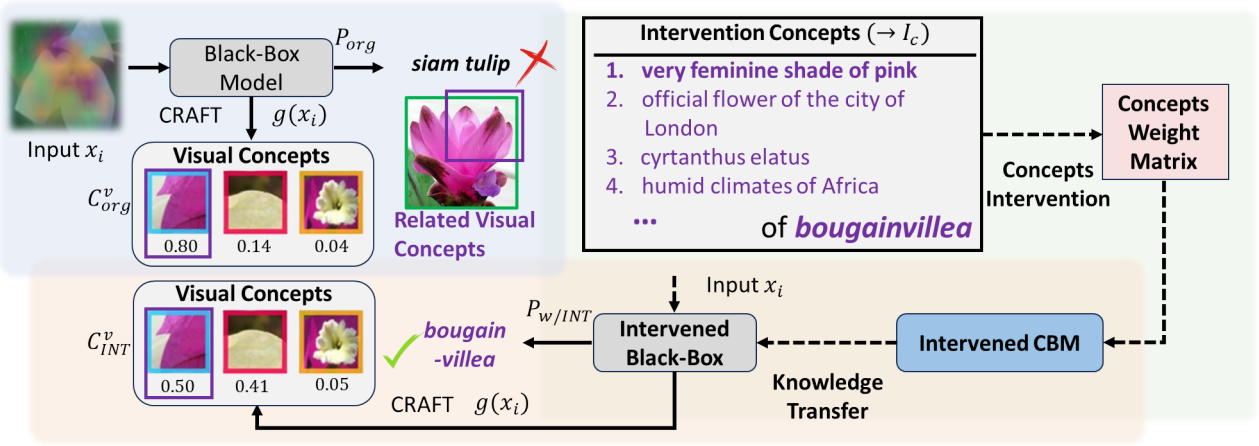

Figure IX. More visualization of BotNet26 on Flower-102.

significantly ($\downarrow 0.06 \rightarrow 0.05$). Combining the misclassifies sample image, we can also see that the visual features corresponding to the intervention concepts are indeed easy to confuse the two classes.

In Figure VII, NFResNet50 misclassifies *BAE 146-300* as *BAE 146-200*. The visually related intervention concept: *"Blue with a grey and white stripe."* and *"Large logo on the tail."* prompts us to delete the related concept about the tail and the color-related feature of aircraft to correct this error. We can find that the second and the third most important visual concept extracted by the original black box before intervention corresponds exactly to the descriptions. After the intervention, we can find that the network relies less on the corresponding color-related feature and tail of the aircraft to classify real class. The concept importance scores of color-related features and aircraft tail dropped from 0.22 to 0.18 and 0.15 to 0.13, respectively.

In Figure VIII, NFResNet50 misclassifies *shiny cowbird* as *American crow*. The visually related intervention concept: *"Black webbed foot."* prompts us to delete the related concept about the foot of bird to correct this error. We can find that the second and the third most important visual concept extracted by the original black box before intervention corresponds exactly to this description. After the intervention, we can find that the network relies less on bird footsteps to classify real class.

In Figure IX, BotNet26 misclassifies *bougainvillea* as *siam tulip*. The visually related intervention concept: *"Very feminine shade of pink."* prompts us to delete the related concept about the pink color of flower to correct this error. We can find that the first most important visual concept extracted by the original black box before intervention corresponds exactly to this description, and the importance score of this concept is 0.80. After the intervention, the black box gives a similar conceptual explanation, but the difference is that the importance score of the intervention concept dropped significantly ($\downarrow 0.30 \rightarrow 0.50$).

In Figure X, BotNet26 misclassifies *Model B200* as *DC-*

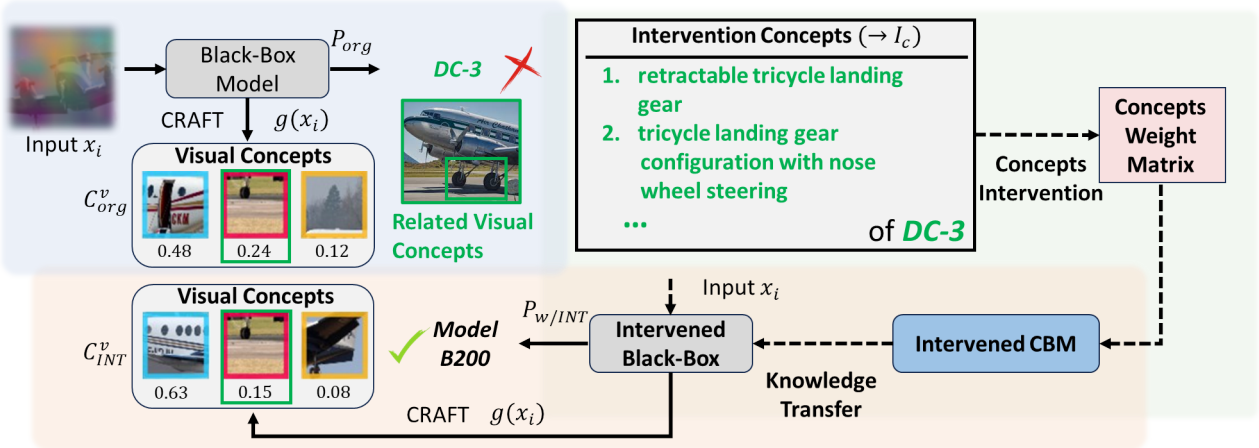

Figure X. More visualization of BotNet26 on FGVC-Aircraft.

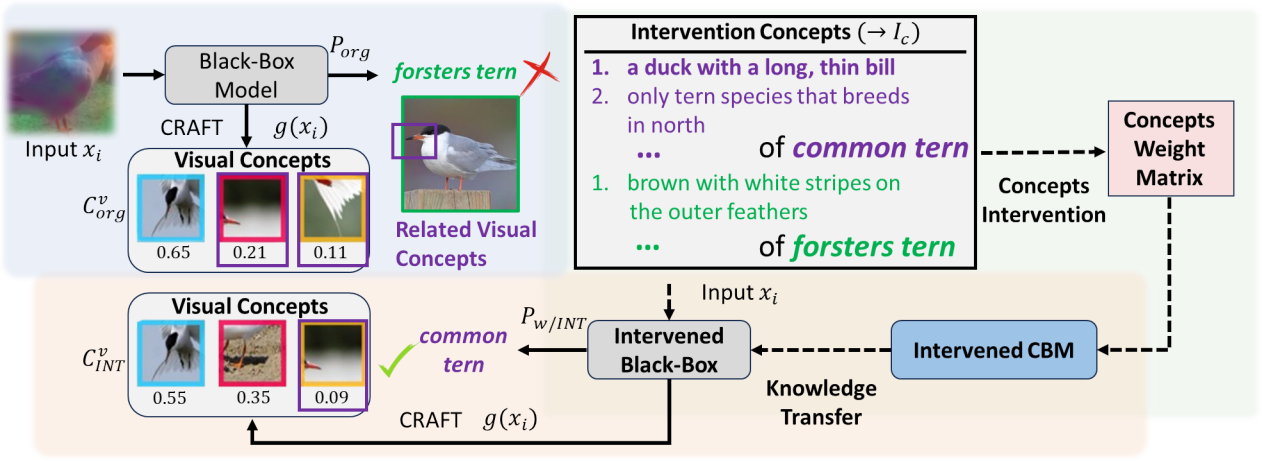

Figure XI. More visualization of BotNet26 on CUB-200.

3. The visually related intervention concept: *"Retractable tricycle landing gear." and "tricycle landing gear configuration with nose wheel steering."* prompts us to delete the related concept about the gear of aircraft to correct this error. We can find that the second most important visual concept extracted by the original black box before intervention corresponds exactly to this description. After the intervention, we can find that the network relies less on gear of the aircraft to classify real class and the original score of the corresponding visual concept dropped from 0.24 to 0.15.

In Figure XI, BotNet26 misclassifies *common tern* as *forsters tern*. The visually related intervention concept: *"A duck with a long, thin bill."* prompts us to delete the related concept about the bill of bird to correct this error. We can find that the second and the third most important visual concept extracted by the original black box before intervention includes this description. After the intervention, we can find that the network relies less on bird bill to classify real class and the original score of the second most important visual concept dropped from 0.21 to 0.09. The third most important visual concept even disappears.

In Figure XII, RexNet100 misclassifies *water lily* as *camellia*. The visually related intervention concept: *"heart-shaped spathe.", "curved spadix." and "prominent spadix in the center of the flower."* prompts us to delete the related concept about the pistil of flower to correct this error. We can find that the second most important visual concept extracted by the original black box before intervention includes corresponding features. However, the stamen feature given in concept 2 can easily be confused between the two classes. After the intervention, we clearly can find that the black box replaced concept 2 with a more representative visual concept of the stamen to *water lily*.

In Figure XIII, RexNet100 misclassifies *MD-87* as *MD-80*. The visually related intervention concept: *"Large t-shaped tail fin.", "white with blue and red stripes.", "large tail fin with the airline's logo.", etc.* prompts us to delete the related concept about the tail with corresponding color-

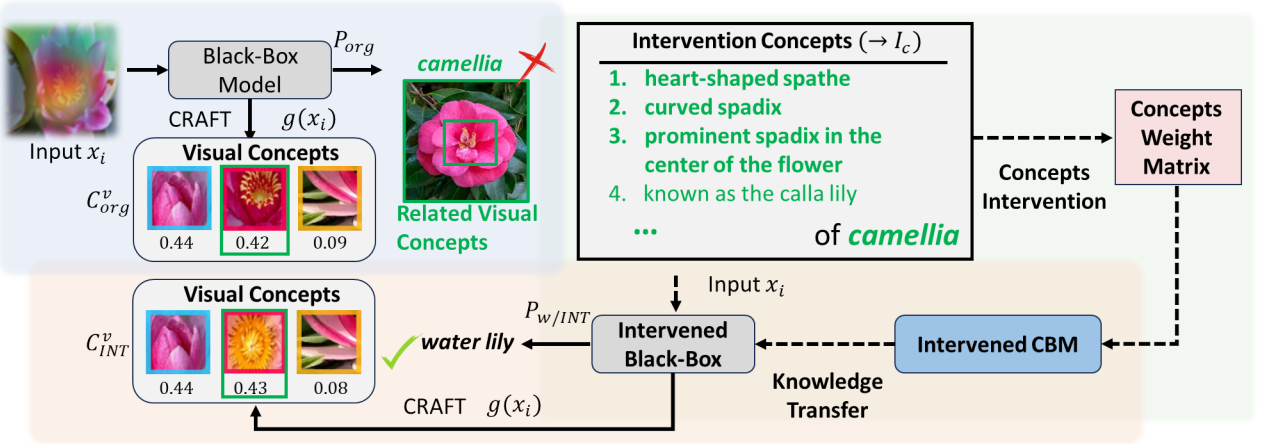

Figure XII. More visualization of RexNet100 on Flower-102.

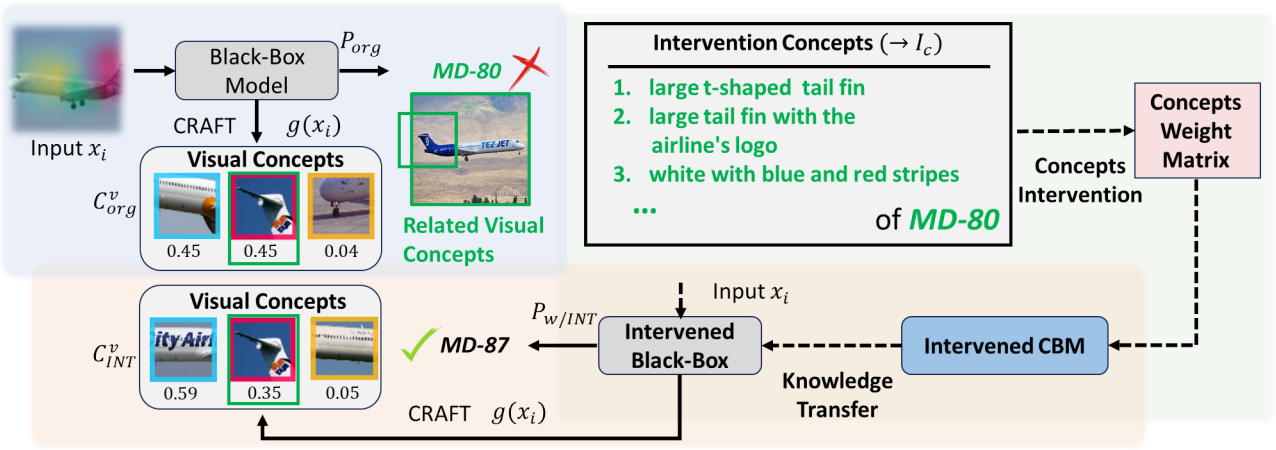

Figure XIII. More visualization of RexNet100 on FGVC-Aircraft.

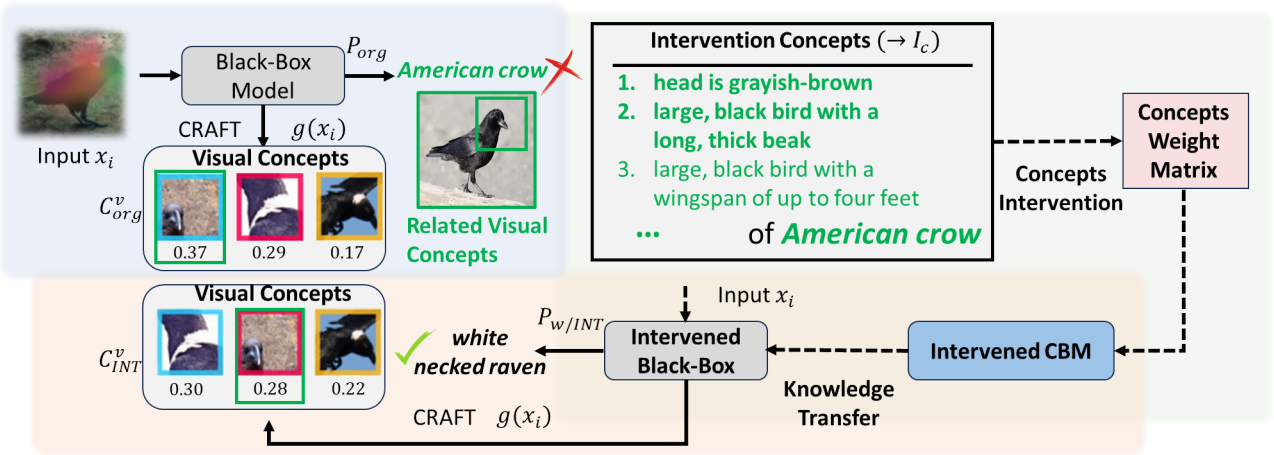

Figure XIV. More visualization of RexNet100 on CUB-200.

related and shape-related feature of aircraft to correct this error. We can find that the second most important visual concept extracted by the original black box before intervention corresponds exactly to this description. After the intervention, we can find that the network relies less on the tail of the aircraft to classify real class and the original score of the corresponding visual concept dropped from 0.45 to 0.35.

In Figure XIV, RexNet100 misclassifies *white necked raven* as *American crow*. The visually related intervention concept: *"Head is grayish-brown." and "large, black bird with a long, thick beak."* prompts us to delete the related concept about the head and beak of bird to correct this error. We can find that the first most important visual concept extracted by the original black box before intervention corresponds exactly to this description. After the intervention, we can find that the network relies less on bird head to classify real class and the original score of the corresponding visual concept dropped from 0.37 to 0.28.